GiCCS: A German in-Context Conversational Similarity Benchmark

Shima Asaadi[1*], Zahra Kolagar[1*], Alina Liebel[1], Alessandra Zarcone[2]

1 Fraunhofer IIS, Germany {shima.asaadi, zahra.kolagar, alina.liebel}@iis.fraunhofer.de
2 Hochschule Augsburg, Germany Alessandra.zarcone@hs-augsburg.de
*** Equal contributions**

Motivation

- Language Models (LMs) in conversational AI
- Semantic Textual Similarity (STS) for the evaluation of LMs
- majority of STS benchmarks: written language resources (non-conversational data), in English
- conversational data and their challenges for STS:
  - more frequent questions and requests
  - similarity based on pragmatic factors triggered by the conversational context

    Could you turn it up a bit? and *I'd like the AC to be colder.*
- limitations of annotation of STS benchmarks using rating scales:
  1. inconsistencies in annotation
  2. scale region bias
  3. fixed granularity issues

We introduce GiCCS, a first German in-context conversational semantic similarity benchmark

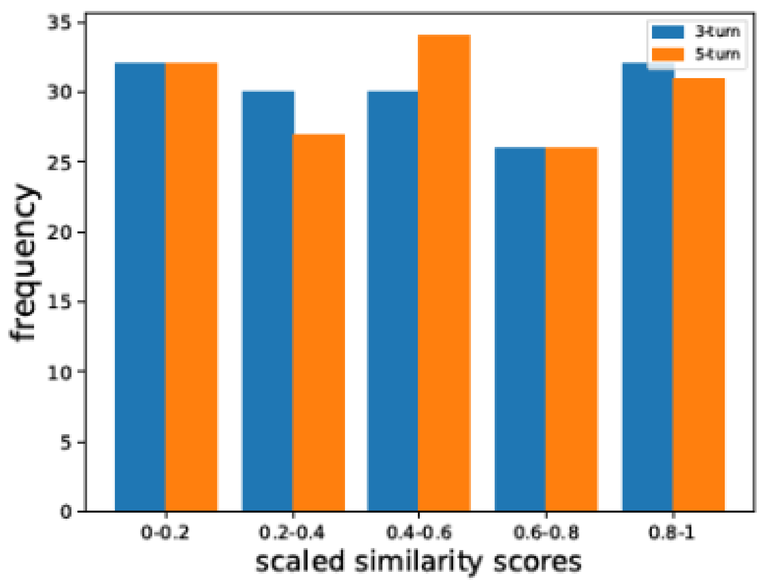
GiCCS includes:

- 300 items
- domain labels
- multi-turn dialogues
- a comparison utterance
- similarity score between the comparison utterance and the last utterance in the dialogue

Annotation Aggregation

- calculate the final semantic similarity scores for dialogue-utterance pairs from the BWS responses
- semantic similarity score of the paired utterance u:

  score(u) = %best(u) - %worst(u)
- similarity scores in [0,1]



Data Collection

Leverage crowdsourced conversational German datasets:

- **CROWDSS** (Frommherz and Zarcone, 2021)
  - contains 113 multi-turn dialogues
  - booking restaurant domain
  - select 24 unique dialogue
  - 12 three-turn and 12 five-turn dialogues
- **BAS SmartKom corpora** (Schiel et al., 2002)
  - select six domains: cinema, fax, navigation, phone, tourist, and tv
  - select 36 unique dialogues
  - 18 three-turn and 18 five-turn dialogues

Create dialogue pairs:

- pair last turn of each dialogue with five hand-written utterances
- utterances were produced by native speakers of German
- pair utterances with different levels of similarity
- obtain 60 dialogues
- each dialogue paired with five sentences
- 300 items in total

GiCCS Lexical Diversity:

Turn	Domain	RTTR	MTLD
3-turn	find_restaurant	3.04	46.22
	find_cinema	1.43	20.34
	find_hotel	1.53	15.90
	find_navigation	1.48	21.31
	find_touristAttraction	1.78	29.60
	find_tvProgram	1.81	30.06
5-turn	find_restaurant	3.01	36.91
	find_cinema	2.08	25.06
	find_hotel	1.53	17.56
	find_navigation	1.62	21.80
	find_touristAttraction	1.76	21.96
	find_tvProgram	2.50	32.03

RTTR: root type-token ratio
MTLD: measure of textual lexical diversity

Inter-Annotator Agreement and Split-Half Reliability Scores:

Dataset	Turn	SHR Spearman	Krippendorff's α	BW-Q	Strong Agreement
BAS SmartKom	3-turn	0.975	0.87	B W	99 100
	5-turn	0.970	0.80	B W	98 100
CROWDSS	3-turn	0.953	0.63	B W	90 94
	5-turn	0.946	0.67	B W	95 97

Strong Agreement: cases where at least four out of five annotators selected the same answer in the best and worst questions

Data Annotation

- Annotation technique: Best-Worst Scaling (BWS)
  - from N = 5 paired utterances in each dialogue generate 2N = 10 distinct 3-tuples
  - obtain 600 distinct 3-tuples to annotate
- Annotation platform: Amazon Mechanical Turk (AMT)
- Annotation task:
  - presented annotators a dialogue at a time, followed by a 3-tuple and asked:
  - *which utterance is most/least similar to the last utterance in the dialogue (best/worst)?*
  - collect five different annotations for each 3-tuple

Experiments: Evaluating LMs

- pairwise STS task
  - predict the cosine similarity score for pairs of utterances
- multiple choice STS task
  - evaluate autoregressive models by considering the dialogue history

Model	Pearson r
distiluse-base-multilingual-cased-v2	0.859
paraphrase-xlm-r-multilingual-v1	0.849
paraphrase-multilingual-MiniLM-L12-v2	0.842
deepset/gbert-large	0.666

Model	Accuracy 3-turn Dialogue	Accuracy 5-turn Dialogue
mGPT	0.133	0.100